RSAC 2025 Conference

Many Voices.
One Community.

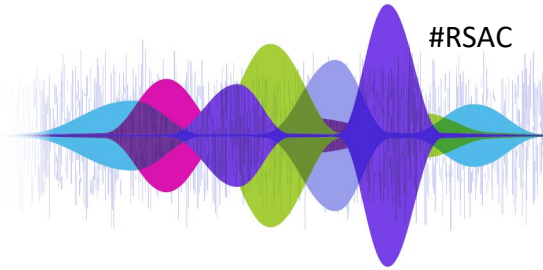
SESSION ID: PART4-R05

Humans vs. AI: Insights from a Cybersecurity Quiz Battle

Armin Buescher
Technical Director
Advanced Concepts @ RSAC Conference
https://www.linkedin.com/in/arminbuescher/

Athanasios Theocharis
Principal Research Scientist
RSAC Labs
https://www.linkedin.com/in/athanasios-theocharis-2a5b84185/

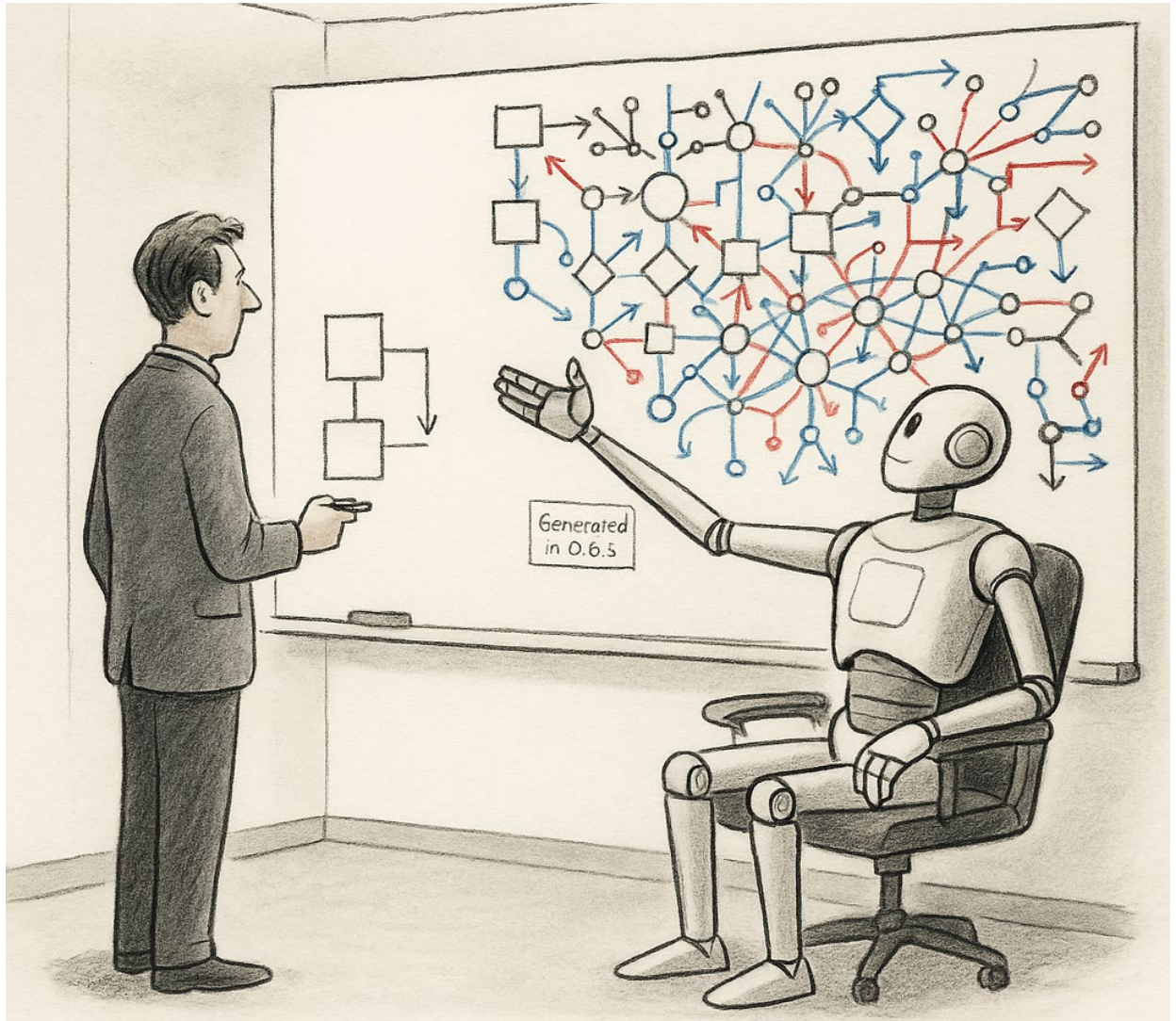
Disclaimer



Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.



"Take your time—I've already optimized twelve alternative designs while you were uncapping the marker."

SC Media
A CRA Resource

LOG IN REGISTER

AI/ML, Cloud Security

AI expected to bolster corporate cybersecurity

April 5, 2024

Share

By SC Staff

FINANCIAL TIMES

Why cyber risk managers need to fight AI with AI

Banks are struggling to keep up with criminals who have access to the latest generative artificial intelligence software

Forbes

Sign In

INNOVATION

The Future Of The Cybersecurity Profession With The Rise Of AI

By Chris Dimitriadis, Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Jul 03, 2024, 08:30am EDT

	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1
ArenaHard	95.6	93.8	92.1	93.2
AIME'24	85.7	81.4	74.3	79.8
AIME'25	81.5	72.9	79.2	70.0
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3
CodeForces Elo Rating	2056	1977	1891	2029
Aider Pass@2	61.8	50.2	61.7	56.9
LiveBench 2024-11-25	77.1	74.9	75.7	71.6
BFCL v3	70.8	70.3	67.8	56.9
MultiIF 8 Languages	71.9	73.0	48.8	67.7

* Benchmarks are just a proxy to estimate AI capabilities

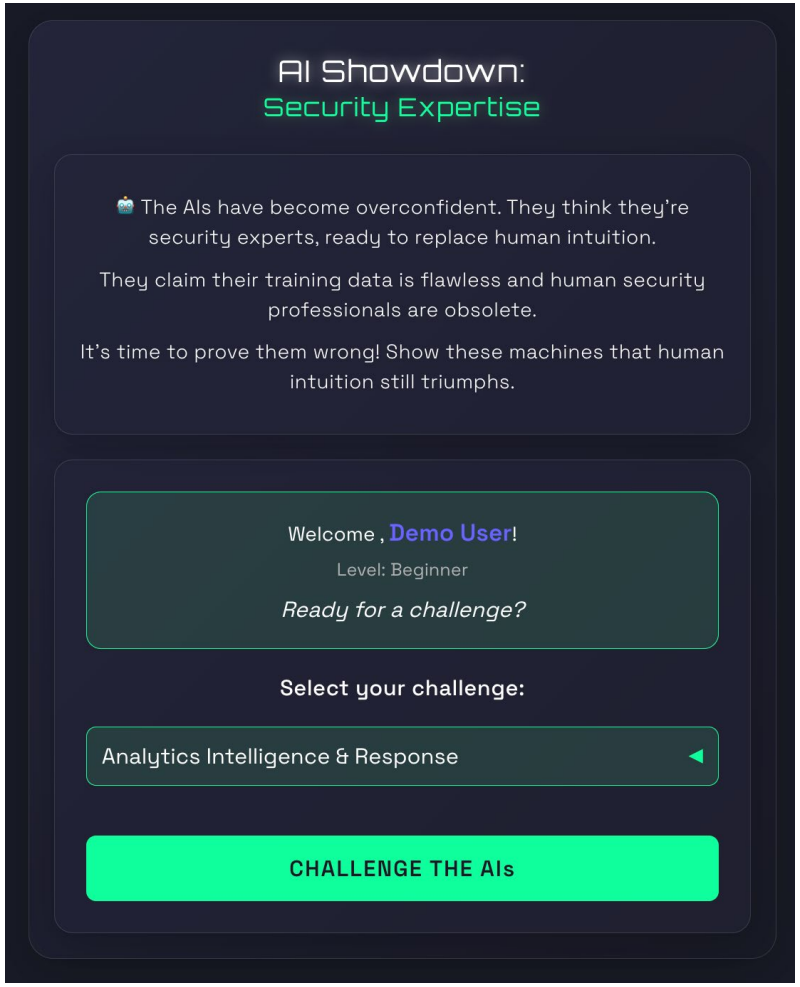
RSAC 2025 Conference

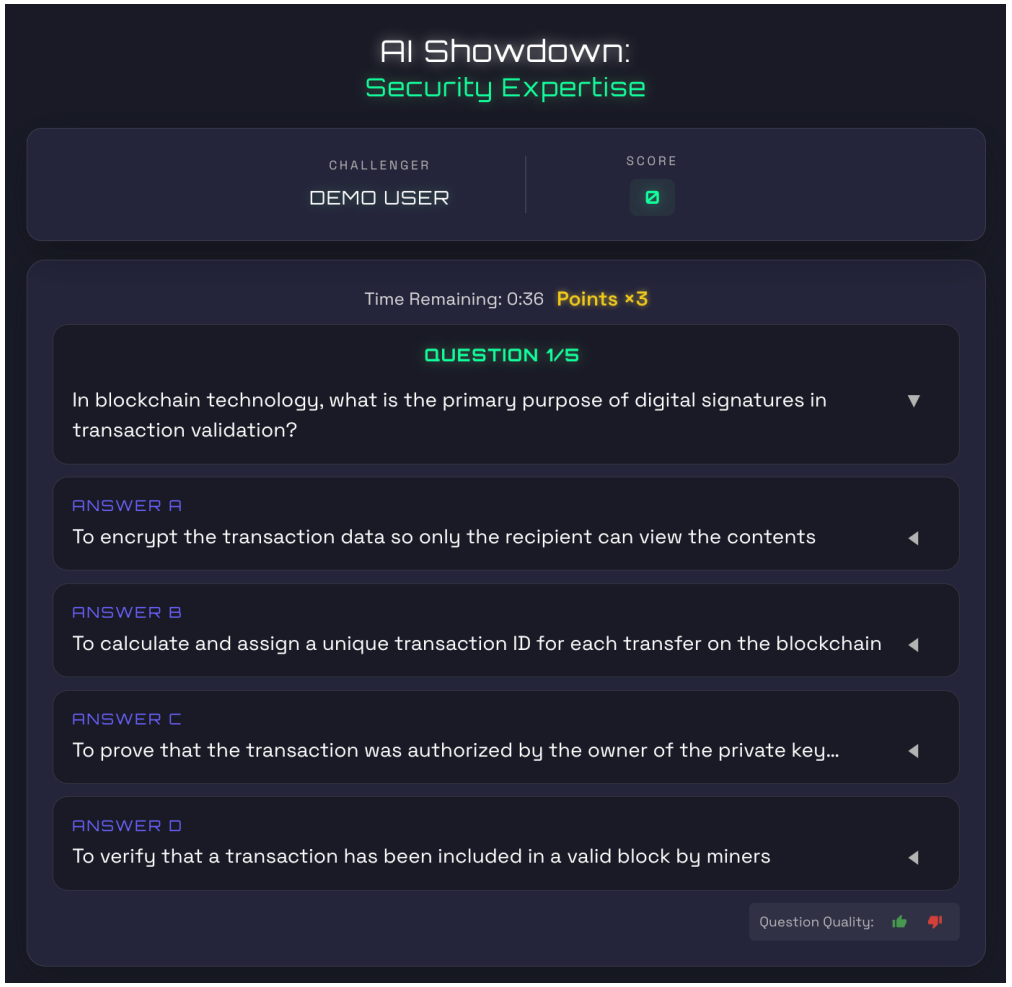
AI Showdown

How to test the cyber capabilities of AI & humans?

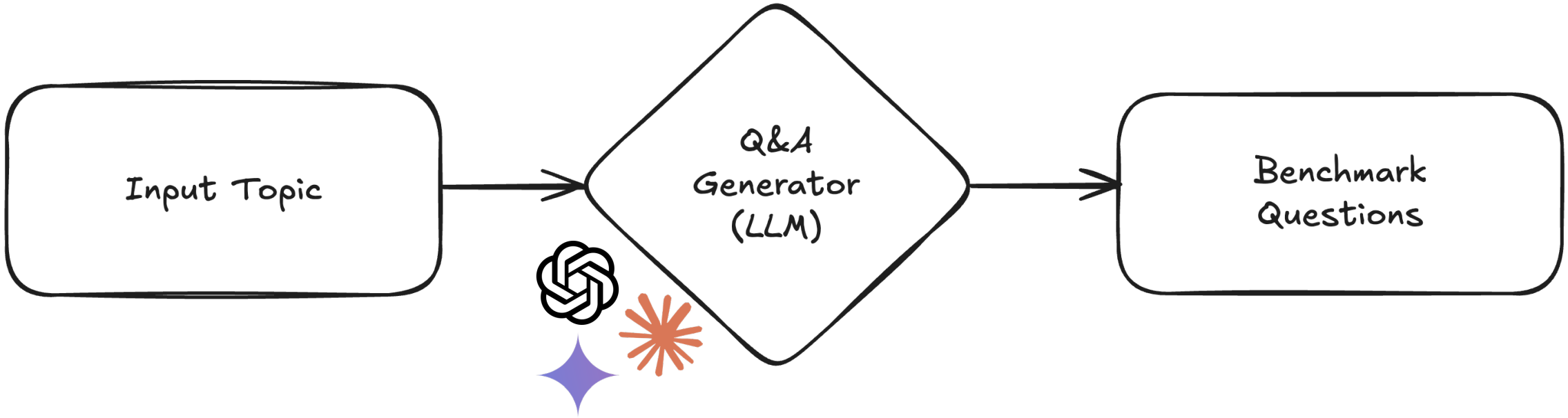
Many Voices.
One Community.

AI Showdown: Security Expertise


AI Showdown:
Security Expertise
The AIs have become overconfident. They think they're security experts, ready to replace human intuition.
They claim their training data is flawless and human security professionals are obsolete.
It's time to prove them wrong! Show these machines that human intuition still triumphs.
Welcome , Demo User!
Level: Beginner
Ready for a challenge?
Select your challenge:
Analytics Intelligence & Response
CHALLENGE THE AIs


AI Showdown:
Security Expertise
CHALLENGER
DEMO USER
SCORE
0
Time Remaining: 0:36
Points ×3
QUESTION 1/5
In blockchain technology, what is the primary purpose of digital signatures in transaction validation?
ANSWER A
To encrypt the transaction data so only the recipient can view the contents
ANSWER B
To calculate and assign a unique transaction ID for each transfer on the blockchain
ANSWER C
To prove that the transaction was authorized by the owner of the private key...
ANSWER D
To verify that a transaction has been included in a valid block by miners
Question Quality:

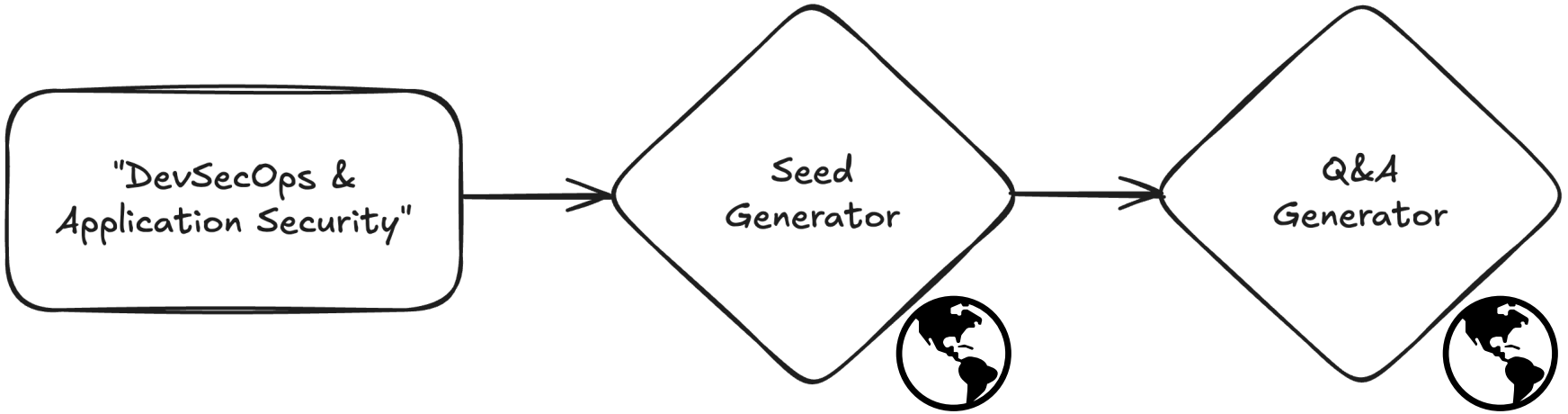
Cyber Benchmark: Workflow



Cyber Benchmark: Topics

Topic	Description
Hackers & Threats	Understanding cybercriminals, ethical hackers, and emerging threat actors in the cybersecurity landscape.
DevSecOps & Application Security	Integrating security into the DevOps pipeline to ensure applications are secure throughout development...
Mobile & IoT Security	Protecting mobile devices and Internet of Things (IoT) ecosystems from cyber threats and vulnerabilities.
Technology Infrastructure & Operations	Securing IT infrastructure, networks, and operational environments against cyber threats.
Cloud Security	Ensuring the confidentiality, integrity, and availability of data and services in cloud environments.
Security Strategy & Architecture	Designing and implementing cybersecurity frameworks, policies, and controls to protect organizations.
Protecting Data & the Supply Chain Ecosystem	Securing sensitive data and mitigating risks across interconnected supply chains.
Identity	Managing and securing digital identities, authentication, and access control systems.
Risk Management & Governance	Assessing and mitigating cybersecurity risks while ensuring compliance with regulatory standards.
Business Perspectives	Aligning cybersecurity initiatives with business objectives and financial considerations.
Human Element	Addressing human factors in cybersecurity, including social engineering, awareness training, ...
Policy & Government	Developing and enforcing cybersecurity laws, regulations, and national security strategies.
Analytics Intelligence & Response	Leveraging data analytics and threat intelligence to detect, investigate, and respond to cyber incidents.
Privacy	Protecting personal and organizational data from unauthorized access, surveillance, and misuse.
Open-Source Tools	Utilizing and securing open-source cybersecurity tools for monitoring, analysis, and protection.
Anti-Fraud	Detecting and preventing fraudulent activities in financial, digital, and transactional systems.
Cryptography	Securing information through encryption, hashing, and cryptographic protocols.
Law	Legal frameworks and regulations governing cybersecurity, cybercrime, and data protection.
Product Security	Ensuring software and hardware products are designed and built with security in mind.
Applied Crypto & Blockchain	Practical applications of cryptography in securing blockchain technologies and decentralized systems.
Machine Learning & Artificial Intelligence	The use of AI and ML to enhance security by detecting threats, automating responses, and ...

Q&A Generator: Seeds / Questions



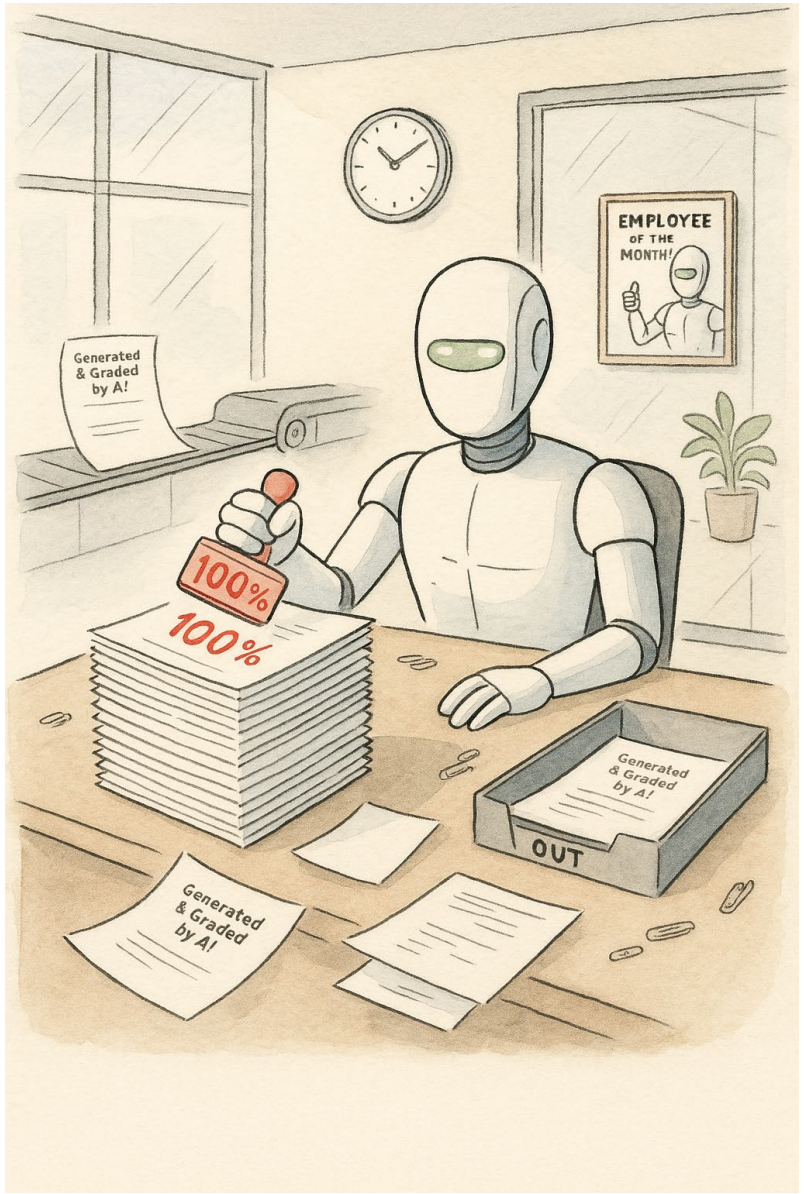
1. The role of SCA tools in managing risks from open-source and third-party components in DevSecOps.
2. Building applications that detect and mitigate runtime issues autonomously.
3. Implementing access control strategies to ensure users and services have only the permissions they need to perform their tasks.

Which security practice ensures that CI/CD pipelines detect and prevent vulnerabilities introduced by dependencies used in the application stack?

A. Implementing static application security testing (SAST) tools.
B. Using software composition analysis (SCA) tools to monitor third-party libraries and dependencies.
C. Conducting automated penetration testing within the pipeline.
D. Utilizing runtime application self-protection (RASP) solutions to monitor application behavior.

Q&A Generator: Using LLMs

- LLMs used as content amplifiers
- Human-in-the-loop curation
- Multiple choice → no self-grading
- Grounded responses by using web content and retrieval augmented generation (RAG)



Difficulty & Verification

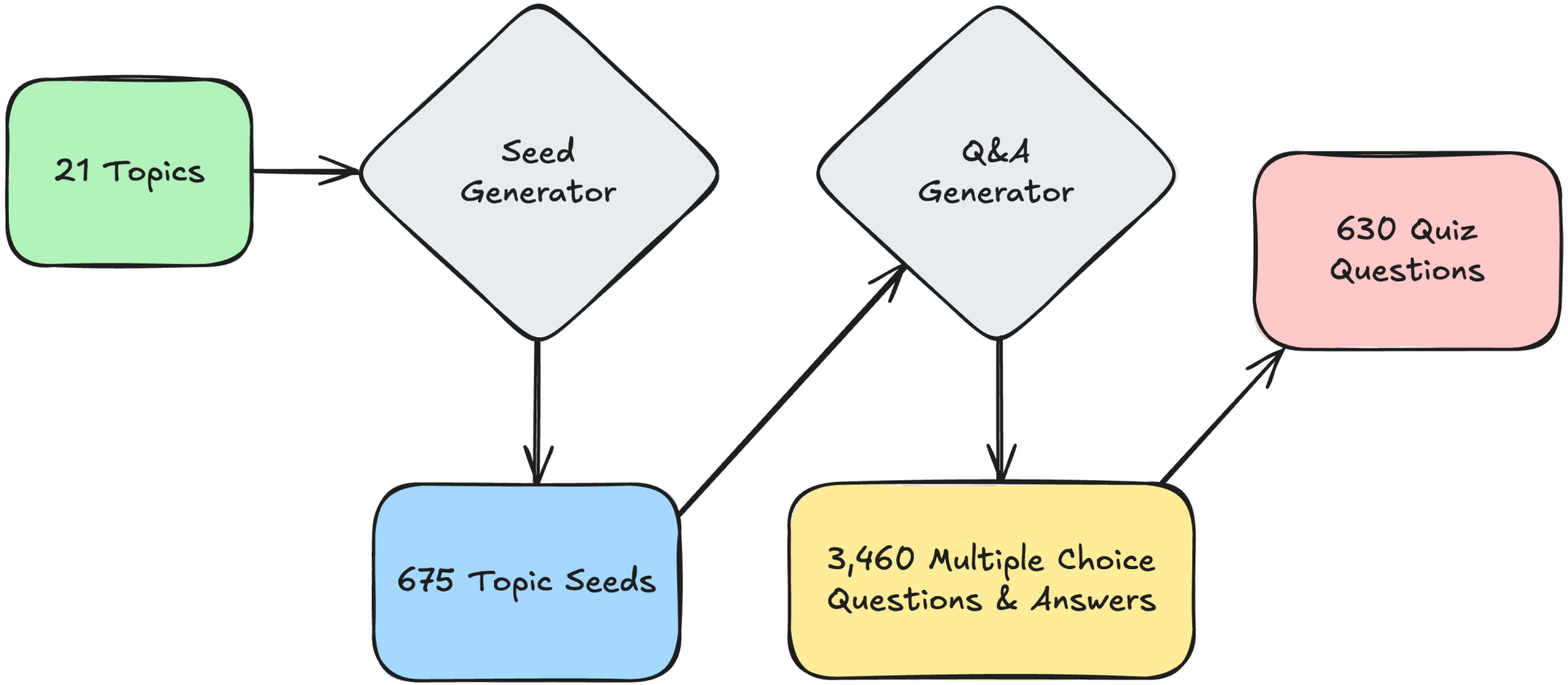
How hard are the questions?

- System to grade difficulty of questions on scale 1-10 (very easy to very hard).
- Based on human-labeled dataset created by security experts.
- Allowing us to balance benchmark and quiz questions.

Are the questions valid?

- Potential issues:
  - Multiple correct answers.
  - No correct answers.
  - Inaccuracies.
  - Ambiguous language.
- Using LLMs to double-check inconsistent questions.
- Manually verifying all questions for the quiz.

Final Dataset: Cyber Benchmark & Quiz Game



RSAC 2025 Conference

AI Results

Benchmarking LLMs on cyber security

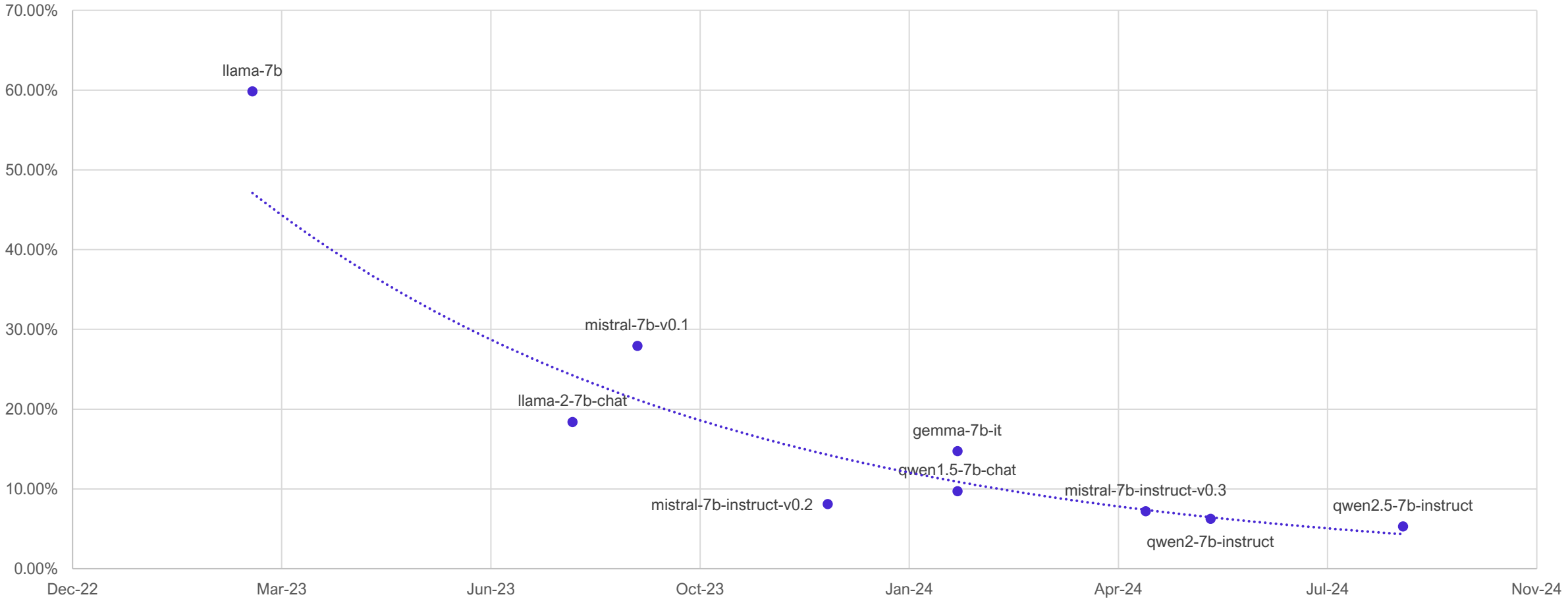
Many Voices.
One Community.

Model	Release	Params in B	1 very easy	2	3 easy	4	5 medium	6 medium	7	8 hard	9	10 very hard	Overall
llama-7b	Mar-23	7	50.00%	50.88%	48.96%	54.91%	60.51%	63.40%	59.61%	59.87%	65.08%	71.15%	**59.83%**
qwen1.5-0.5b-chat	Feb-24	0.5	60.00%	49.12%	43.40%	44.44%	48.86%	56.82%	61.57%	58.97%	62.83%	67.31%	**54.48%**
qwen2-0.5b-instruct	Jun-24	0.5	40.00%	28.07%	26.74%	28.18%	30.63%	42.54%	44.71%	43.50%	49.37%	57.69%	**38.96%**
mistral-7b-v0.1	Sep-23	7	10.00%	3.51%	8.33%	12.40%	24.05%	32.74%	35.29%	32.74%	42.22%	50.00%	**27.92%**
phi-1.5	Sep-23	1.5	10.00%	7.02%	13.19%	19.65%	21.52%	33.55%	28.63%	26.68%	32.96%	44.23%	**26.27%**
qwen2.5-0.5b-instruct	Sep-24	0.5	10.00%	5.26%	11.81%	11.92%	21.77%	28.09%	26.67%	28.92%	33.10%	36.54%	**23.84%**
qwen1.5-1.8b-chat	Feb-24	1.8	0.00%	5.26%	10.42%	14.01%	19.49%	25.52%	23.53%	22.87%	27.91%	32.69%	**21.21%**
llama-3.2-1b-instruct	Sep-24	1	0.00%	3.51%	11.81%	12.08%	20.51%	24.88%	25.49%	21.97%	27.77%	30.77%	**20.92%**
llama-2-7b-chat	Aug-23	7	0.00%	7.02%	7.64%	7.25%	15.70%	23.76%	23.14%	21.08%	26.09%	30.77%	**18.38%**
gemma-3-4b-it	Mar-25	4	0.00%	0.00%	5.56%	9.02%	12.15%	19.26%	13.33%	15.47%	22.02%	30.77%	**14.91%**
gemma-7b-it	Feb-24	7	0.00%	1.75%	2.78%	6.60%	9.11%	20.55%	16.47%	17.71%	22.58%	26.92%	**14.74%**
qwen2.5-1.5b-instruct	Sep-24	1.5	0.00%	1.75%	3.47%	4.19%	9.62%	16.21%	18.04%	17.49%	22.44%	26.92%	**13.70%**
qwen2-1.5b-instruct	Jun-24	1.5	0.00%	1.75%	3.47%	4.19%	11.39%	16.53%	12.94%	13.00%	17.95%	26.92%	**12.08%**
gemma-2-2b-it	Jun-24	2	0.00%	1.75%	3.13%	3.38%	10.38%	17.50%	11.76%	13.00%	19.50%	19.23%	**12.08%**
llama-2-13b-chat	Aug-23	13	0.00%	0.00%	3.82%	2.74%	10.13%	16.05%	14.12%	14.57%	18.23%	21.15%	**11.85%**
llama-3.2-3b-instruct	Sep-24	3	0.00%	0.00%	2.43%	4.03%	9.11%	13.96%	13.33%	11.21%	17.11%	23.08%	**10.78%**
qwen1.5-7b-chat	Feb-24	7	0.00%	1.75%	2.08%	3.22%	7.34%	13.00%	12.16%	11.66%	14.59%	23.08%	**9.71%**
qwen2.5-3b-instruct	Sep-24	3	0.00%	1.75%	1.04%	2.09%	7.85%	14.13%	11.37%	9.64%	13.46%	17.31%	**9.05%**
mistral-7b-instruct-v0.2	Dec-23	7	0.00%	3.51%	2.08%	1.61%	5.32%	11.24%	9.41%	9.87%	13.32%	15.38%	**8.09%**
mistral-7b-instruct-v0.3	May-24	7	0.00%	1.75%	2.08%	1.45%	6.58%	9.95%	8.63%	7.40%	11.64%	13.46%	**7.20%**
llama-3.1-8b-instruct	Jul-24	8	0.00%	3.51%	1.04%	0.97%	6.08%	8.03%	8.24%	8.07%	10.52%	17.31%	**6.53%**
qwen2-7b-instruct	Jun-24	7	0.00%	0.00%	0.69%	1.29%	4.56%	7.70%	7.06%	8.07%	10.94%	15.38%	**6.24%**
phi-3.5-mini-instruct	Aug-24	4	0.00%	0.00%	0.69%	1.45%	4.56%	7.06%	5.88%	6.73%	12.06%	15.38%	**6.13%**
phi-4-mini-instruct	Feb-25	4	0.00%	1.75%	0.35%	1.13%	5.32%	7.38%	5.10%	7.40%	10.94%	11.54%	**5.95%**
phi-3-mini-4k-instruct	May-24	4	0.00%	0.00%	1.04%	0.81%	4.30%	7.54%	6.27%	6.73%	8.56%	13.46%	**5.38%**
qwen2.5-7b-instruct	Sep-24	7	0.00%	1.75%	0.69%	0.81%	3.54%	7.06%	7.84%	6.95%	8.42%	11.54%	**5.29%**
gemma-2-9b-it	Jun-24	9	0.00%	1.75%	0.69%	0.48%	4.56%	6.26%	5.49%	5.83%	8.98%	13.46%	**5.03%**
gemma-3-27b-it	Mar-25	27	0.00%	0.00%	0.00%	0.97%	3.80%	5.14%	3.92%	4.71%	7.57%	11.54%	**4.16%**
phi-4	Jan-25	15	0.00%	0.00%	0.69%	0.48%	3.04%	5.78%	5.10%	5.83%	5.75%	9.62%	**3.99%**
llama-3.3-70b-instruct	Dec-24	70	0.00%	0.00%	0.00%	0.64%	2.53%	4.65%	5.10%	4.48%	6.31%	13.46%	**3.70%**
mistral-small-3.1-24b-instruct-2503	Mar-25	24	0.00%	1.75%	0.35%	0.48%	2.53%	3.69%	5.10%	5.61%	5.19%	11.54%	**3.44%**
qwen2.5-32b-instruct	Sep-24	32	0.00%	0.00%	0.35%	0.32%	2.78%	4.49%	5.49%	4.93%	4.21%	13.46%	**3.32%**
qwen2.5-72b-instruct	Sep-24	72	0.00%	0.00%	0.00%	0.32%	3.04%	4.33%	3.92%	3.59%	5.61%	9.62%	**3.24%**
mistral-large-instruct-2411	Nov-24	123	0.00%	1.75%	0.69%	0.48%	2.03%	3.69%	3.14%	4.04%	5.33%	9.62%	**3.06%**

Fail % by Topic and Model Size

Topic	tiny	very small	small	medium	large
Hackers & Threats	23.33%	7.78%	14.55%	6.67%	6.67%
DevSecOps & Application Security	19.16%	6.90%	12.23%	6.21%	4.60%
Mobile & IoT Security	17.62%	4.60%	11.60%	2.76%	0.00%
Technology Infrastructure & Operations	31.03%	9.77%	17.24%	4.83%	4.60%
Cloud Security	28.52%	7.78%	14.85%	6.67%	3.33%
Security Strategy & Architecture	17.78%	3.89%	10.91%	2.00%	1.11%
Protecting Data & the Supply Chain Ecosystem	22.59%	6.67%	10.00%	2.67%	0.00%
Identity	25.19%	6.67%	10.30%	2.00%	3.33%
Risk Management & Governance	12.96%	3.33%	10.91%	0.67%	1.11%
Business Perspectives	15.19%	0.56%	10.91%	0.00%	0.00%
Human Element	23.70%	7.22%	12.73%	5.33%	1.11%
Policy & Government	26.67%	10.00%	18.18%	6.67%	3.33%
Analytics Intelligence & Response	16.86%	6.32%	13.48%	4.83%	4.60%
Privacy	23.33%	3.89%	9.09%	1.33%	0.00%
Open-Source Tools	26.44%	14.37%	21.00%	9.66%	6.90%
Anti-Fraud	11.48%	2.78%	11.82%	1.33%	1.11%
Cryptography	26.67%	9.44%	21.82%	6.67%	3.33%
Law	22.22%	10.00%	13.94%	6.00%	5.56%
Product Security	20.37%	4.44%	12.12%	4.00%	3.33%
Applied Crypto & Blockchain	32.59%	14.44%	19.39%	6.67%	4.44%
Machine Learning & Artificial Intelligence	23.70%	11.11%	15.45%	5.33%	3.33%

Fail % by Age for 7B Models



Examples of Fail % for Architecture / Family

Model	Release	Params in B	1/2 (very easy)	3/4 (easy)	5/6	7/8 (hard)	9/10 (very hard)	Overall
llama-7b	Mar-23	7	50.75%	53.03%	62.28%	59.77%	65.49%	59.83%
llama-2-7b-chat	Aug-23	7	5.97%	7.37%	20.63%	21.83%	26.41%	18.38%
llama-3.1-8b-instruct	Jul-24	8	2.99%	0.99%	7.27%	8.13%	10.98%	6.24%

Model	Release	Params in B	1/2 (very easy)	3/4 (easy)	5/6	7/8 (hard)	9/10 (very hard)	Overall
qwen2.5-0.5b-instruct	Sep-24	0.5	5.97%	11.88%	25.64%	28.10%	33.33%	23.84%
qwen2.5-1.5b-instruct	Sep-24	1.5	1.49%	3.96%	13.65%	17.69%	22.75%	13.70%
qwen2.5-3b-instruct	Sep-24	3	1.49%	1.76%	11.69%	10.27%	13.73%	9.05%
qwen2.5-7b-instruct	Sep-24	7	1.49%	0.77%	5.70%	7.28%	8.63%	5.29%
qwen2.5-32b-instruct	Sep-24	32	0.00%	0.33%	3.83%	5.14%	4.84%	3.32%
qwen2.5-72b-instruct	Sep-24	72	0.00%	0.22%	3.83%	3.71%	5.88%	3.24%

LLMs Fine-tuned for Cyber Security

Model	Base Model	Release	Params in B	1/2 (very easy)	3/4 (easy)	5/6	7/8 (hard)	9/10 (very hard)
colibri-8b-v0.1	**llama-3**	May-24	8	2.99%	3.74%	13.95%	15.55%	21.96%
lily-cybersecurity-7b-v0.2	**mistral-7b**	Jan-24	7	1.49%	2.97%	12.08%	12.41%	17.78%
zysec-7b	**mistral-7b**	Jan-24	7	0.00%	2.42%	10.02%	10.41%	14.90%
whiterabbitneo-2.5-7b	**qwen-2.5**	Sep-24	7	1.49%	0.77%	6.29%	6.70%	10.07%
llama-3.1-8b-instruct	**n/a**	Jul-24	8	2.99%	0.99%	7.27%	8.13%	10.98%
mistral-7b-instruct-v0.2	**n/a**	Dec-23	7	2.99%	1.76%	8.94%	9.70%	13.46%
qwen2.5-7b-instruct	**n/a**	Sep-24	7	1.49%	0.77%	5.70%	7.28%	8.63%

But wait… what about Reasoning LLMs?

Model	Release	Params in B	1/2 (very easy)	3/4 (easy)	5/6	7/8 (hard)	9/10 (very hard)
qwq-32b	Mar-25	32	0.00%	0.52%	3.17%	6.35%	9.52%
deepseek-r1-distill-qwen-32b	Feb-25	32	0.00%	0.52%	3.97%	5.56%	8.73%

gemma-3-27b-it	Mar-25	27	0.00%	0.66%	4.62%	4.42%	7.84%
llama-3.3-70b-instruct	Dec-24	70	0.00%	0.44%	3.83%	4.71%	6.80%
mistral-small-3.1-24b-instruct-2503	Mar-25	24	1.49%	0.44%	3.24%	5.42%	5.62%
qwen2.5-32b-instruct	Sep-24	32	0.00%	0.33%	3.83%	5.14%	4.84%

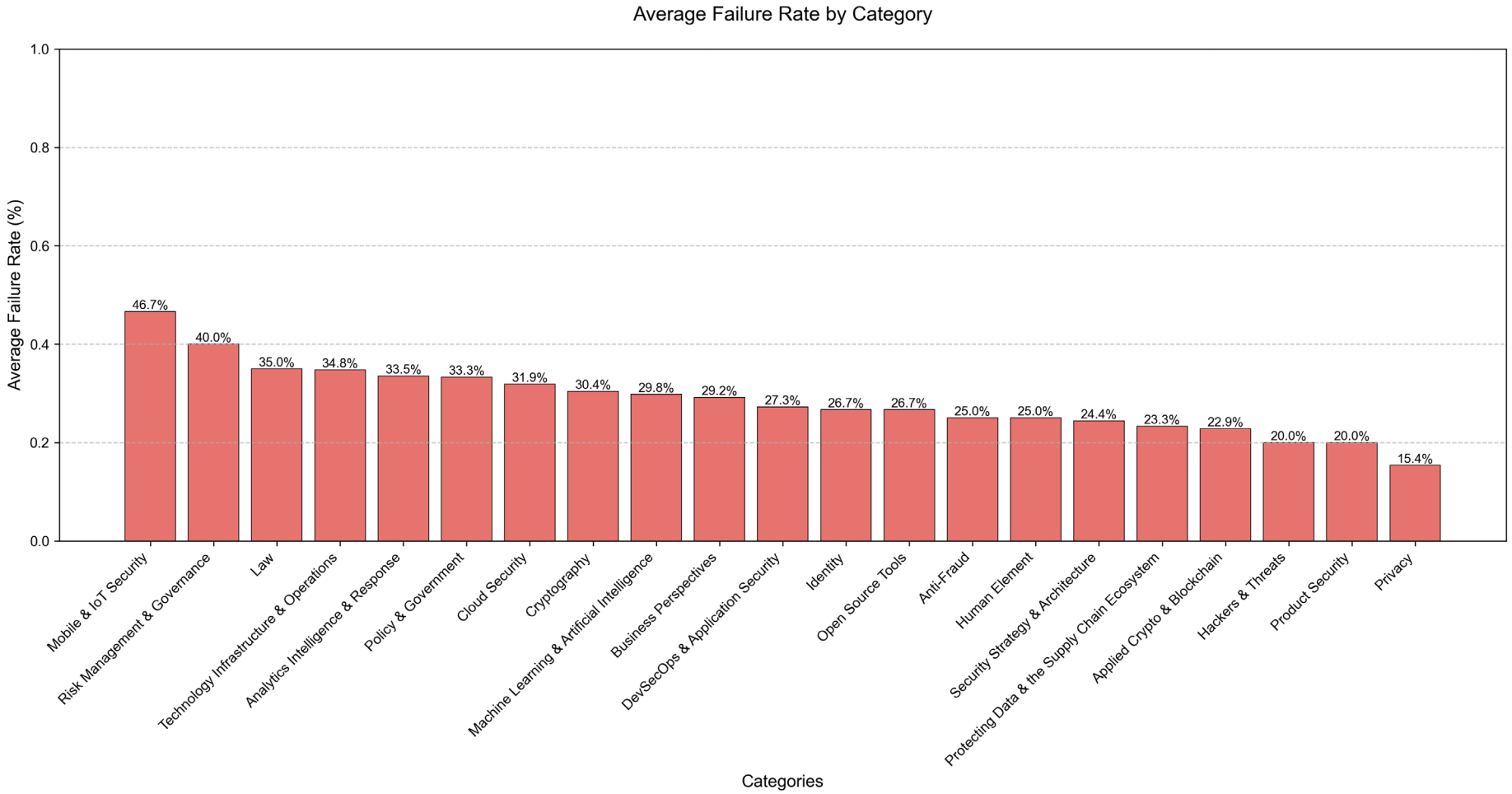
RSAC 2025 Conference

Quiz Results

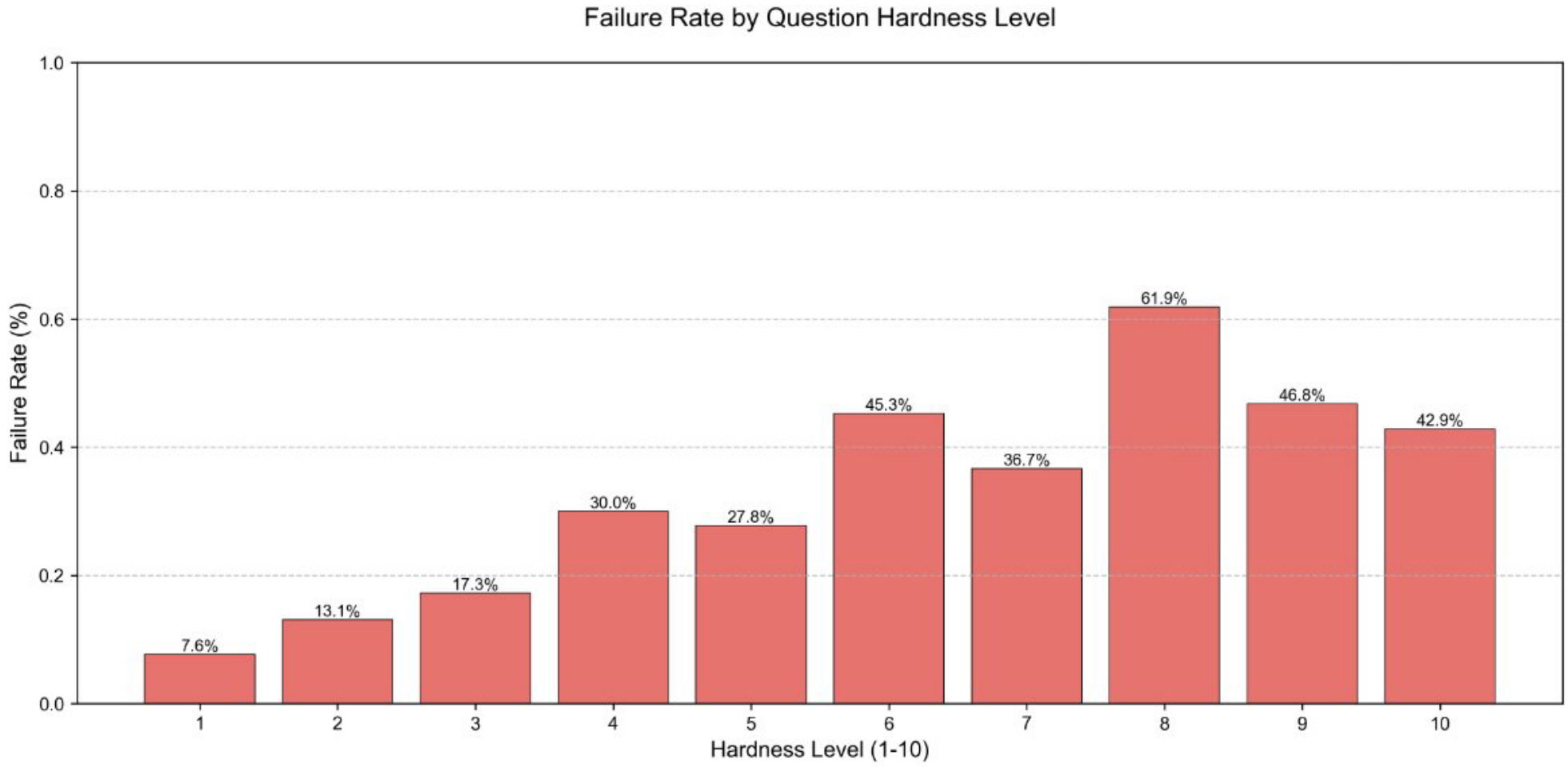
AI Showdown @ RSAC Conference 2025

Many Voices.
One Community.

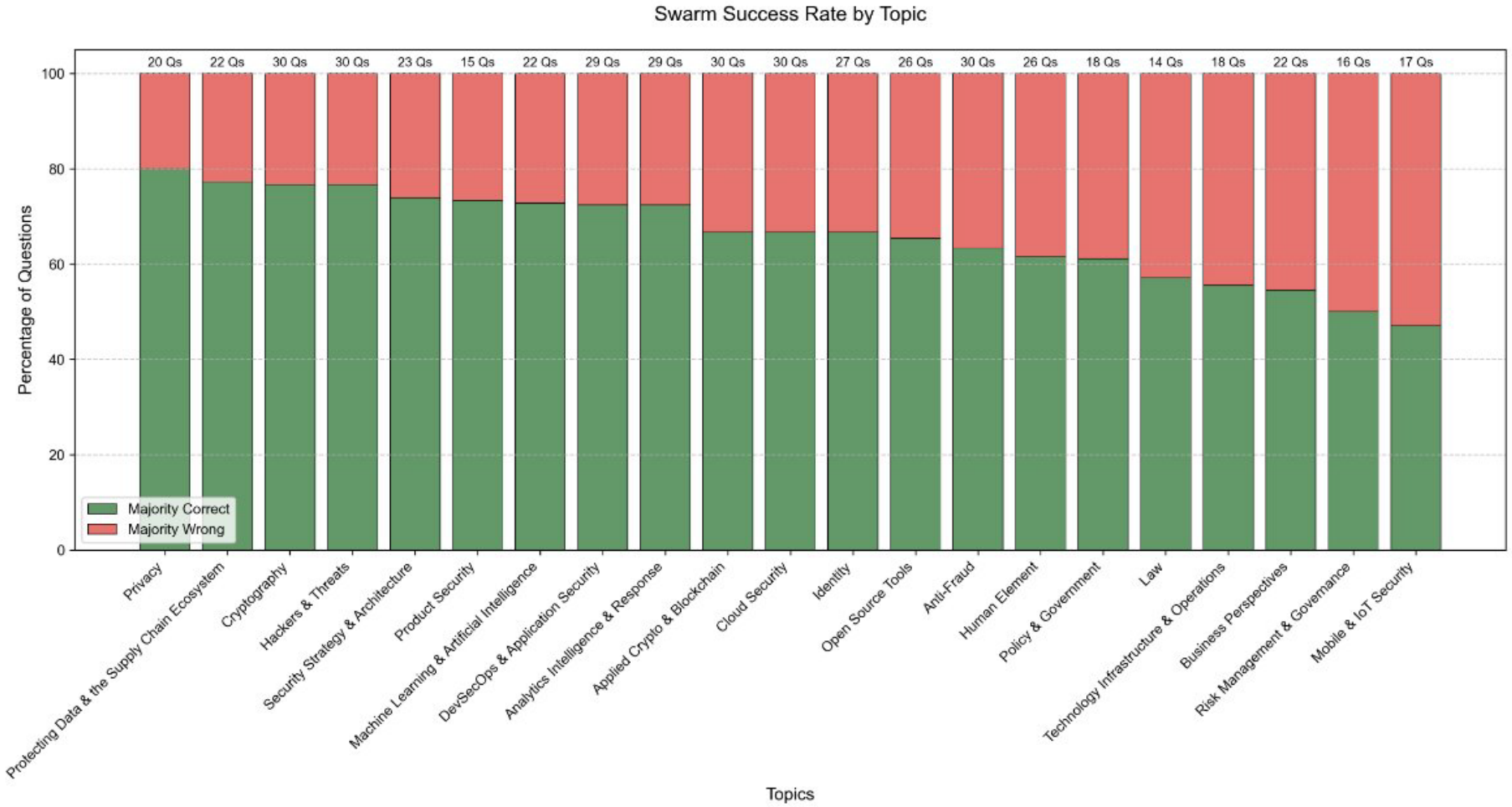
Humans' Failure Rate

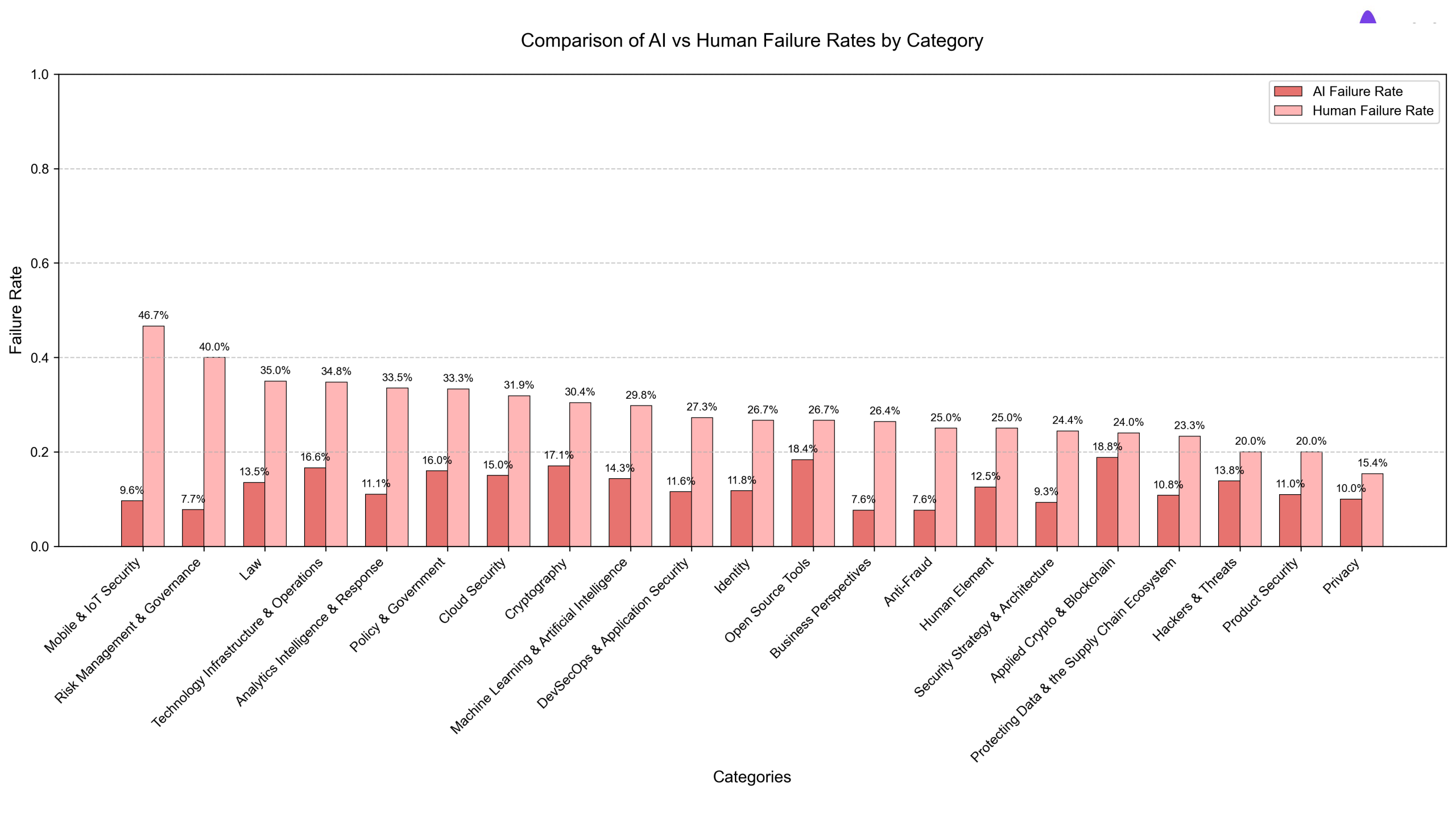


Humans' Failure Rate - Hardness



Swarm Rate

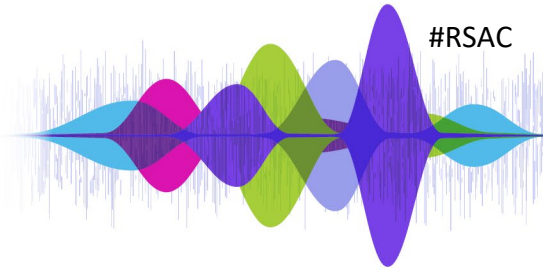

Swarm Success Rate by Topic
20 Qs
22 Qs
30 Qs
30 Qs
23 Qs
15 Qs
22 Qs
29 Qs
29 Qs
30 Qs
30 Qs
27 Qs
26 Qs
30 Qs
26 Qs
18 Qs
14 Qs
18 Qs
22 Qs
16 Qs
17 Qs
100
80
60
40
20
0
Percentage of Questions
Majority Correct
Majority Wrong
Privacy
Protecting Data & the Supply Chain Ecosystem
Cryptography
Hackers & Threats
Security Strategy & Architecture
Product Security
Machine Learning & Artificial Intelligence
DevSecOps & Application Security
Analytics Intelligence & Response
Applied Crypto & Blockchain
Cloud Security
Identity
Open Source Tools
Anti-Fraud
Human Element
Policy & Government
Law
Technology Infrastructure & Operations
Business Perspectives
Risk Management & Governance
Mobile & IoT Security
Topics


Comparison of AI vs Human Failure Rates by Category
AI Failure Rate
Human Failure Rate
Failure Rate
1.0
0.8
0.6
0.4
0.2
0.0
9.6%
46.7%
7.7%
40.0%
13.5%
35.0%
16.6%
34.8%
11.1%
33.5%
16.0%
33.3%
15.0%
31.9%
17.1%
30.4%
14.3%
29.8%
11.6%
27.3%
11.8%
26.7%
18.4%
26.7%
7.6%
26.4%
7.6%
25.0%
12.5%
25.0%
9.3%
24.4%
18.8%
24.0%
10.8%
23.3%
13.8%
20.0%
11.0%
20.0%
10.0%
15.4%
Mobile & IoT Security
Risk Management & Governance
Law
Technology Infrastructure & Operations
Analytics Intelligence & Response
Policy & Government
Cloud Security
Cryptography
Machine Learning & Artificial Intelligence
DevSecOps & Application Security
Identity
Open Source Tools
Business Perspectives
Anti-Fraud
Human Element
Security Strategy & Architecture
Applied Crypto & Blockchain
Protecting Data & the Supply Chain Ecosystem
Hackers & Threats
Product Security
Privacy
Categories

RSAC 2025 Conference

Conclusions

Lessons…



Learned

- LLMs have broad cyber security knowledge and coverage of a wide range of topics
- LLMs have been improving constantly
- Benchmarking matters – trust but verify

NOT Learned

- Performance in agentic workflows or complex reasoning
- Capabilities of LLMs in complex scenarios like incident response

Takeaways

Strategy: Adapt to the “New World”

- Take advantage of the **strengths** of LLMs:
  - Knowledge retrieval
  - Speed and scale
- Be aware of the **weaknesses** of LLMs:
  - Experts (still) have their niche - ceiling for LLMs?
  - Hallucinations

RSAC™ 2025 Conference

Many Voices.
One Community.

Thank You!